Interface pour le calcul de risques dans les maladies génétiques

Application au cancer du sein avec le modèle de Claus-Easton

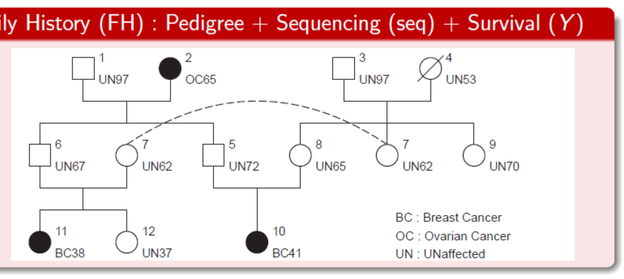
A. Lefebvre[1] A. de Pauw[2] O. Bouaziz[3] M. Le Mentec[2] A. Dupré[2]
D. Stoppa-Lyonnet[2] G. Nuel[1]

Context of the breast cancer

- 1st cancer in women. Affects 54,000 women in France each year
- Complex disease due to an accumulation of mutations in oncogenes and/or supressor of tumor genes (BRCA 1/2, PALB2, RAD51, etc.)
- **Inherited mutation** in 10 to 15% of the cases leading more often to severe family histories (FH)

Data structure in genetic diseases modeling

Family History (FH) : Pedigree + Sequencing (seq) + Survival (Y)



Variant data (var)
- Known variant or VUS (Uncertain Significance)
- Database
- Functional tests
- Molecular dynamics

Cancer pathology (patho)
- MSI status (Lynch) (MicroSatellite Instability)
- BRAF mutation (Lynch)
- ER, PR, HER2 status (Breast)
- invasive/in situ (Breast)

V : the set of variant status
var : the set of variant data
X : the set of all possible values for the genotypes
seq : the set of sequencing data
Y the set of phenotypes (survival data) } → FH = $\{seq, Y\}$
patho : the set of pathology reports

$$\mathbb{P}(\text{var}, V, X, \text{seq}, Y, \text{patho}) = \prod_j \mathbb{P}(\text{var}_j|V_j)\mathbb{P}(V_j) \times \prod_{i, Y_i \neq \text{UN}} \mathbb{P}(\text{patho}_i|X_i) \times \prod_i \mathbb{P}(X_i|X_{\text{pat}_i}, X_{\text{mat}_i})\mathbb{P}(\text{seq}_i|X_i)\mathbb{P}(Y_i|X_i)$$

compute ⇒ posterior **variant status** **carrier risk** **tumoral risk**

The Claus-Easton model

- Developped by Claus et al. (1991) and Easton et al. (1993)
- Used in first intention at the Curie Institute
- Focuses on the genotypes (X) and the phenotypes (Y)
- Autosomal, biallelic, dominant mode of inheritance
- Estimated allele frequency ($f = 0.33\%$); hazard functions per genotype (λ_0 and λ_1) derived from Easton's estimated densities

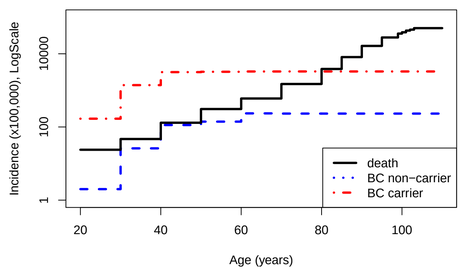


Figure 1: Annual death incidences in the French female population (INED, 2017) and annual breast cancer incidences for non-carriers and carriers estimated from Easton et al. (1993).

Our objectives: **Implement** the Claus-Easton model in a Bayesian network (**sum-product algorithm**) combined with **survival data** and develop a user-friendly **interface**.

Implementation

$$\mathbb{P}(X, Y) = \prod_i \underbrace{\mathbb{P}(X_i|X_{\text{pat}_i}, X_{\text{mat}_i})}_{\text{genotypes}} \underbrace{\mathbb{P}(Y_i|X_i)}_{\text{phenotypes}}$$

The genotypes (unobserved) : $X = (X_i)_{i=1,\dots,n} \in \{00, 01, 10, 11\}^n$

- Mode of inheritence and allele frequency from the **Claus-Easton literature**
- **Hardy-Weinberg** for the founders (assumption)
- **Mendelian transmission** for the offsprings (assumption)

The phenotypes (observed (FH))

$$Y_i(\text{survival data}) = \begin{cases} \{T_i > \tau_i\} & \text{if } i \text{ is censored (UN) at age } \tau_i \\ \{T_i = \tau_i\} & \text{if } i \text{ is affected (BC or OC) at age } \tau_i \end{cases}$$

with T_i being the age at disease onset for individual i.

- **Incidence**
$$\lambda(t) = \begin{cases} \lambda_0(t) & \text{for } X = 00 \text{ (non-carrier (NC))} \\ \lambda_1(t) & \text{for } X \neq 00 \text{ (carrier (C))} \end{cases}$$
- **Survival functions of** T_i :
$$\begin{cases} S_0(t) = \exp\left(-\int_0^t \lambda_0(t)\right) dt & \text{for NC} \\ S_1(t) = \exp\left(-\int_0^t \lambda_1(t)\right) dt & \text{for C} \end{cases}$$
- **Conditional probabilities :**

- For a **censored** individual at age τ_i : $\mathbb{P}(Y_i|X_i) = \begin{cases} S_0(\tau_i) & \text{for NC} \\ S_1(\tau_i) & \text{for C} \end{cases}$
- For an **affected** individual at age τ_i : $\mathbb{P}(Y_i|X_i) = \begin{cases} S_0(\tau_i)\lambda_0(\tau_i) & \text{for NC} \\ S_1(\tau_i)\lambda_1(\tau_i) & \text{for C} \end{cases}$

Computation of the posterior carrier probability

Problematic
We denote by $\mathrm{K}_i(X_i, X_{\text{pat}_i}, X_{\text{mat}_i})$, the potential related to i. $\mathrm{K}_i(X_i, X_{\text{pat}_i}, X_{\text{mat}_i}) = \sum_{Y_i \in \text{FH}} \mathbb{P}(X_i|X_{\text{pat}_i}, X_{\text{mat}_i})\mathbb{P}(Y_i|X_i)$

Using the **Bayes rule**, for any individual j,

$$\mathbb{P}(X_j = x_j|\text{FH}) = \frac{\mathbb{P}(X_j = x_j, \text{FH})}{\mathbb{P}(\text{FH})} = \frac{\sum_{X \setminus x_j} \prod_i \mathrm{K}_i(X_i, X_{\text{pat}_i}, X_{\text{mat}_i})}{\sum_X \prod_i \mathrm{K}_i(X_i, X_{\text{pat}_i}, X_{\text{mat}_i})}$$

With $X \in \{00, 10, 01, 11\}^n \rightarrow 4^n$configurations

The sum-product algorithm (Koller and Friedman, 2009) is equivalent to the latest version of Elston-Stewart algorithm (Totir et al., 2009)

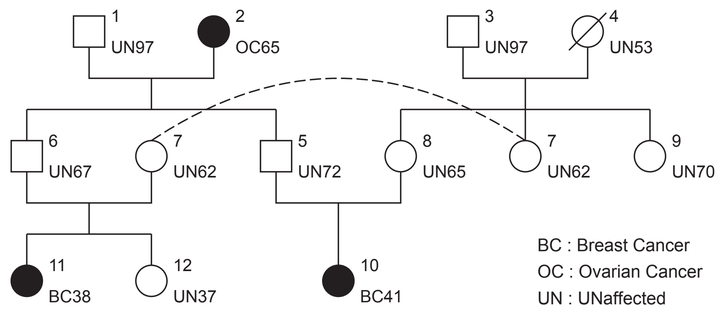


Figure 2: Hypothetical family with a severe FH.

Using the conditional independencies between the individuals and the minimum fill-in heuristic, we obtain the following junction tree :

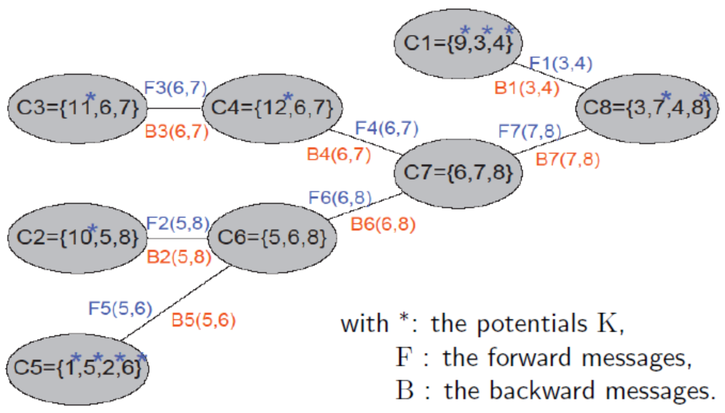


Figure 3: One junction-tree obtained from our hypothetical family.

e.g. $\mathrm{F}_3(6,7) = \sum_{X_{11}} \mathrm{K}_{11}(X_{11}, X_6, X_7)$

$\mathrm{F}_4(6,7) = \sum_{X_{12}} \mathrm{K}_{12}(X_{12}, X_6, X_7)\mathrm{F}_3(6,7)$ $\quad \mathrm{B}_4(6,7) = \sum_{X_8} \mathrm{F}_6(6,8)\mathrm{B}_7(7,8)$

All F&B messages computed once for any later marginal or joint distribution needed ⇒ Complexity $\mathcal{O}(4^n) \rightarrow \mathcal{O}(n \times 4^k)$
k:tree-width =3 in general (2 parents & 1 child in a clique),
4 in case of a loop (mating loop, consanguinity),
5 or more exceptionally (several loops).

e.g.: $\mathbb{P}(X_6, X_7, X_8|\text{FH}) \propto \mathrm{F}_4(6,7)\mathrm{F}_6(6,8)\mathrm{B}_7(7,8)$

$\mathbb{P}(X_6|\text{FH}) \propto \sum_{X_8} \mathrm{F}_6(6,8)\mathrm{B}_6(6,8)$

Results : The interface

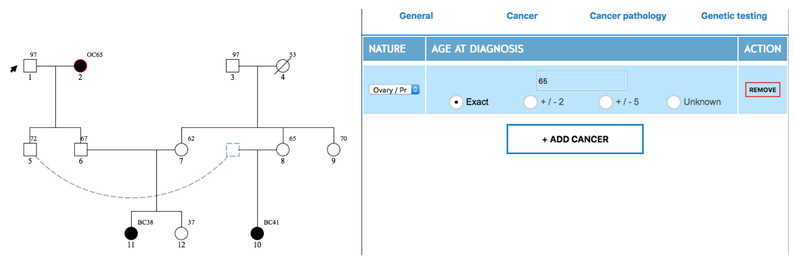


Figure 4: Example of a data entry of a hypothetical family history with the interactive interface.

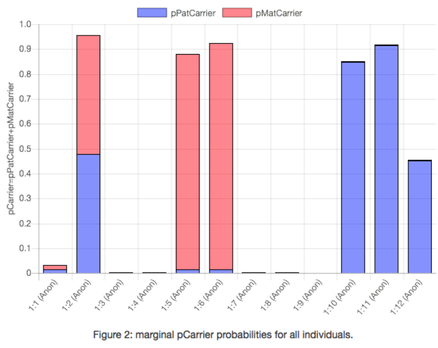


Figure 2: marginal pCarrier probabilities for all individuals.

Ind	pPatCarrier	pMatCarrier	pCarrier
1:1 (Anon) M 97	0.0160167	0.0160167	0.0320333
1:2 (Anon) F OC65	0.47883	0.47883	0.95766
1:3 (Anon) M 97	0.00258042	0.00258042	0.00516084
1:4 (Anon) F 53	0.00142826	0.00142826	0.00285651
1:5 (Anon) M 72	0.0153911	0.865806	0.881197
1:6 (Anon) M 67	0.0160017	0.908384	0.924385
1:7 (Anon) F 62	0.00308557	0.00170409	0.00478965

Figure 5: Probability of being a carrier in a barplot (on top) and in a table (at the bottom) obtained with the interface for each individual of our hypothetical family.

Results - Carrier predisposition

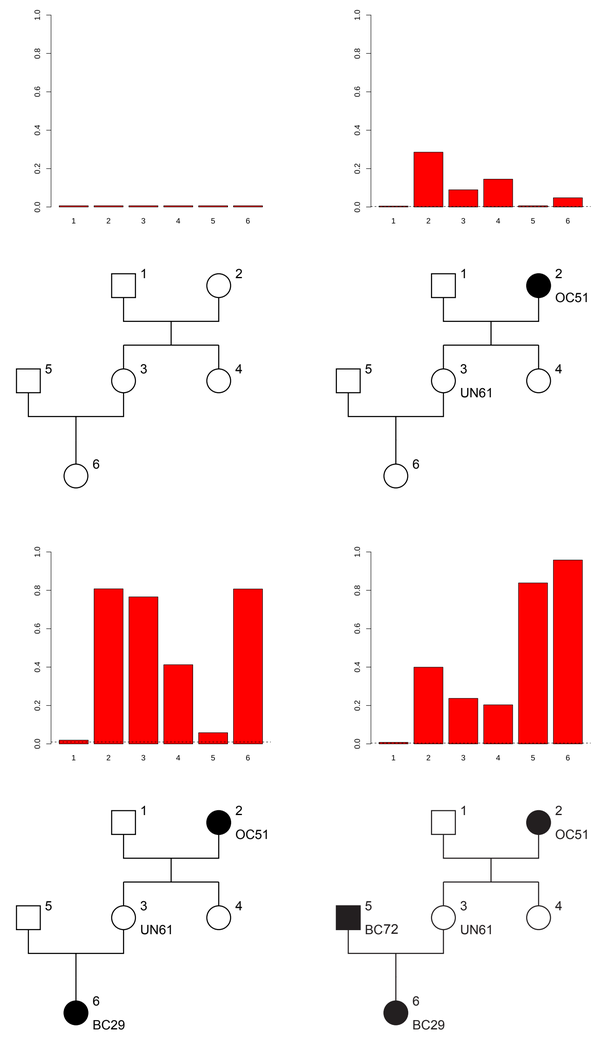


Figure 6: Probability of being a carrier for each individual of a family with evolving FH.

Computation of the disease risks prediction

We denote by $\pi(\tau) = \mathbb{P}(X_i \neq 00(\text{carrier})|\text{FH})$

- With no competing risk of death
$\mathbb{P}(T \leq t|\text{FH}) = 1 - S(t|\text{FH})$ with
$S(t|\text{FH}) = \sum_{X_i} \mathbb{P}(T > t, X_i|\text{FH}) = \pi(\tau)\frac{S_1(t)}{S_1(\tau)} + (1 - \pi(\tau))\frac{S_0(t)}{S_0(\tau)}$
- Quantities needed to compute the competing risk
$\pi(t|\text{FH}) = \frac{\pi(\tau)S_1(t)}{S(t|\text{FH})S_1(\tau)}$
$\boldsymbol{\lambda_{\text{disease}}(t|\text{FH})} = \pi(t|\text{FH})\lambda_1(t) + (1 - \pi(t|\text{FH}))\lambda_0(t)$
- With competing risk of death
$T^* = \min(T_{\text{disease}}, T_{\text{death}})$
$\lambda_{\text{both}}(t|\text{FH}) = \boldsymbol{\lambda_{\text{disease}}(t|\text{FH})} + \lambda_{\text{death}}(t)$
$\mathbb{P}(T \leq t|\text{FH}) = \int_\tau^t S_{\text{both}}(u)\boldsymbol{\lambda_{\text{disease}}(u)}\boldsymbol{du}$

Results - Tumoral risk

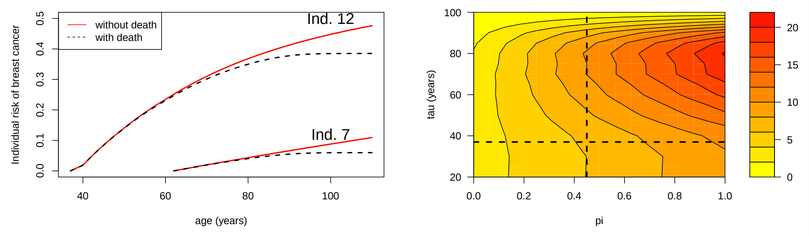


Figure 7: Left-panel : Disease risk for indivudals 12 and 7 in our hypothetical family. The risk is computed with and without the consideration of the competing risk of death. Right-panel : Percentage of error made (difference) while computing the disease risk up to the age 100 without vs with taking into account the competing risk of death for different couples (π, τ). The dashed lines represent the error made for individual 12 in our hypothetical family.

References

Elisabeth B Claus, N Risch, and W Douglas Thompson. Genetic analysis of breast cancer in the cancer and steroid hormone study. *American journal of human genetics*, 48(2):232, 1991.

DF Easton, DT Bishop, D Ford, and GP Crockford. Genetic linkage analysis in familial breast and ovarian cancer: results from 214 families. the breast cancer linkage consortium. *American journal of human genetics*, 52(4):678, 1993.

INED. Death incidence in France, period 2012-2014. `https://www.ined.fr/en/everything_about_population/data/france/deaths-causes-mortality/mortality-tables/`, 2017.

Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Liviu R Totir, Rohan L Fernando, and Joseph Abraham. An efficient algorithm to compute marginal posterior genotype probabilities for every member of a pedigree with loops. *Genetics Selection Evolution*, 41(1):52, 2009.

Acknowledgment

This work has been funded by Institut Curie (M2 internship) and *Ligue Nationale Contre le Cancer* (PhD grant).

[correspondance: `lefebalex@gmail.com` or `nuel@math.cnrs.fr`]

[1] Probability and Statistics (LPSM, CNRS 8001), Sorbonne Université, Paris; [2] Genetic Oncology (INSERM U830), Institut Curie, Paris; [3] Applied Mathematics (MAP5, CNRS 8145), Université Paris Descartes, France.